Hybrid Knowledge Model to Help End-Users Retreive Relevant Information

Reena Shetty, Pierre-Michel Riccio, Joël Quinqueton

Hybrid Knowledge Model to Help End-Users Retrieve Relevant Information

Reena T. N. Shetty	Pierre-Michel Riccio	Joël Quinqueton
PhD Student – ENSMP, Paris,	Assistant Professor,	Professor LIRMM,
LGI2P Laboratory, EMA SITE EERIE, Nîmes	LGI2P Laboratory, EMA SITE EERIE, Nîmes	University Montpellier – II, Montpellier
reena.shetty@ema.fr	pierre-michel.riccio@ema.fr	jq@lirmm.fr

Abstract

In this short paper, we present the knowledge representation model called Extended Semantic Network. The basic idea of this proposal is to imagine data representation techniques which can reason beyond the classical techniques in information retrieval systems. It is argued that by employing hybrid techniques one can address the good recall with powerful deduction problems. Our objective here is to achieve semi-supervised knowledge representation technique with good accuracy and minimum human intervention, using the heuristically developed information processing and integration methods.

1 Introduction

The recent years have seen data and knowledge bases progressively converge towards the electronic media, attributed to the ever mounting use of the World Wide Web (WWW). For many people, the World Wide Web has become an essential means of providing [Katz, 1997] and searching for information leading to large amount of data accumulation. Searching web in its present form is however an infuriating experience for the fact that the data available is both superfluous and diverse in form. In the past, one has seen several propositions supporting Web users find relevant information to their queries. Most of these retrieval methods have limited their abilities to specific tasks.

One of the most commonly used methods in information retrieval is the document retrieval technique based on keyword search. Document retrieval is commonly defined as the matching of stated user query term with the useful parts of free text records or reproduction of records. These records could be of any type mainly unstructured text, such as bibliographic records, newspaper records or just paragraphs in an instruction handbook. User's queries can normally vary from multi-sentence full descriptions about information needed to a few words. Vast majority of retrieval systems currently in use range from simple Boolean systems to systems using statistical or natural language processing techniques.

In such retrieval models, users submit their queries corresponding to the information they desire to be made available. The existing retrieval systems return voluminous sets of documents as the query result. Thus it is the sole responsibility of the users to find the relevant information to their query within the returned result set, consequentially investing more time in analyzing the output results due to its immenseness. Moreover, many of the results found turn out to be extraneous and one can find some of the more important links not being listed in the result set.

The logical deduction for such under performing situation is the facts that, majority of the existing data resources in its present form are designed for human comprehension. When using these data with machines, it becomes highly impracticable to obtain good results without human interventions at regular levels. So, one of the major challenges faced by the consumers of web era is to imagine intelligent tools and theories in knowledge representation and processing techniques which can support and enable efficient analyzing of data by machines.

2 Objectives

In this section we will give a brief overview on Question Answering (QA) techniques and how our approach can be useful in addressing the requirements of the QA techniques. We will then in the following sections describe the different modules of our approach in more detail. Then the expert evaluated results of our model are discussed and presented in the result section. This is followed by a conclusion and future perspectives.

One of the most promising researches carried out in this field is that of Question answering [Moldovan, 2002] nature of information retrieval system. This technique generally enables retrieval of answers posed to questions in natural language on a given collection of documents. The technique deals with a wide range of question types ranging from fact, list, definition, hypothetical, semantically-constrained to cross-lingual questions. Here, the research collections targeted normally vary from small local collections of documents to large sets internal documents.

Users here pose questions in natural languages [Voorhees, 2002] and the system analyses their query and returns answers to the queries in the structure of list of short and specific answers. This often proves more useful in case of specific information needs and also greatly reduces the analyzing time. This ability makes it to be regarded as the next

step beyond classical search engines. The objective here is to intelligently query data representations to obtain the most relevant information as answers.

This requires knowledge representation (KR) [Sowa, 2000] models with shrewd data representation. These models should necessarily cover vast knowledge bases simply because the main principle for the QA to work efficiently is to have access to a good search corpus. Research in knowledge representation was initially centered on formalisms that were characteristically tuned to deal with relatively smaller knowledge representation techniques but on contrary provide efficient deductions.

In the current scenario, for a modern knowledge representation model to be useful in realistic applications, it is important that it handles large data sets and provide strong query deductions. The two most important features that a QA looks for in any model is its

- ability to cover vast knowledge bases and
- phrasing redundant data in different ways under different contexts.

3 Our Approach

Observing these necessities in QA approach we propose a novel knowledge representation model called the extended semantic network. The basic idea of extended semantic network is to identify an efficient knowledge representation method to overcome the existing constraints in efficiently identifying the right information in data retrieval systems. Here, we try to realize a model that can cover vast knowledge bases by equally retaining its high inference capabilities.

To realize this goal we put our ideas into practice via a three phase approach. The first phase consists in processing large amount of textual information using mathematical models to make our proposal scalable. The second phase involves in manually constructing small semantic networks based on our model derived from KL-ONE [Brachman and Schmolze, 1985]. The third and the last phase consist in examining carefully and efficiently the various possibilities of integrating information obtained from our mathematical information model with that of the manually developed mind deduction model.

The manual model is constructed by initially identify concepts representing each domain with the help of domain experts and then these concepts are retransmitted to the experts with the relational links to be used in interconnecting these concepts based on their relations. These links are provided with a weight value as an additional parameter. This value will be considered as the cost of passing over the chosen path between concepts. The links used in the semantic network are detailed in later sections. This model forms the concept network with high precision which will guide us in organising our recall model obtained from the word network obtained from the automated mathematical model where edge values are designated using the results of the mathematical models applied on the documents.

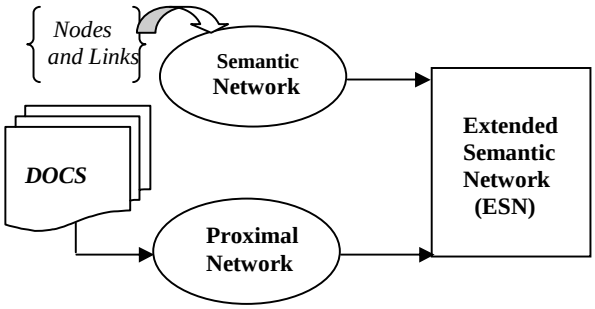

Figure 1: Schematic representation of extended semantic network

The idea here is to develop an innovative approach obtained by combination of features from man and machine theory of concept [Sowa, 1984], whose results can be of enormous use in the latest knowledge representation, classification, retrieval, pattern matching and ontology development research fields. This paper highlights the methods we employ in information processing and integration for visualising a novel knowledge representation [Quillian, 1968] method to be used in information retrieval techniques.

3.1 Proximal Network

The basic theory of proximity is concerned with the arrangement or categorisation of entities that relate to one another often believed to favour interactive learning, knowledge creation and innovation. When a number of entities are close in proximity a relationship is implied and if entities are logically positioned; they connect to form a structural hierarchy. Our proximal network model is built based on this structural hierarchy, of word proximity in documents.

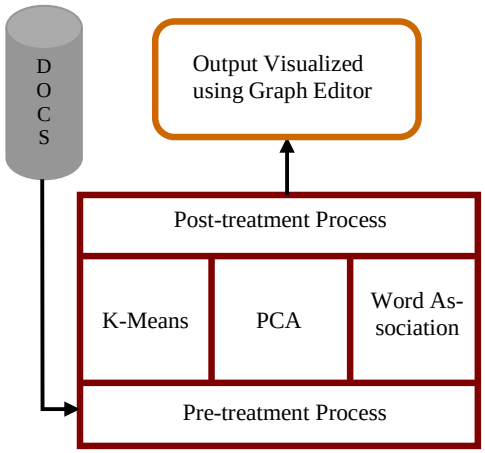

Figure 2: Block diagram representing proximal network prototype

This approach is largely employed to enable processing of large amount of data [Winston et al., 1987] in a considerably small time. Unlike the NLP models our model does not depend on heavy calculations but on contrary use simple statistical and clustering logics to obtain the results equally efficient but by consuming much lesser time and resource. Another important aspect of this approach is its ability to automatically process the input data into a network of concepts interconnected with mathematically established relations forming a recall focused approach.

The proximal network model involves three processing phase, firstly the pre-treatment process where the documents related are analyzed in 2 stages and an output of word document matrix is obtained. This matrix is then passed on to the intermediate process consisting of the 3 processing agents and is analyzed by the data mining and clustering algorithms namely K-means clustering, Principle component analysis and Word association to obtain an output of word pair matrix with a value between each word pair. This value is the proximity between the word pair in the projected space depending on their occurrence in the contents of the documents processed [Mahé et al., 2001].

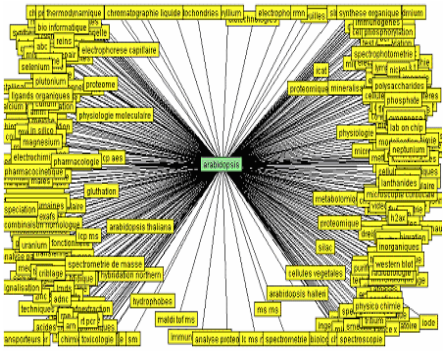

Figure 3: An extract of proximal network representation for arabidopsis using graph editor

This is further subjected to post-treatment process where partial stemming is carried on the word pair matrix depending on case based requirement. The output is then stored into a Mysql database and visualized using the Graph Editor, a java application developed by us for visualization and easy editing of networks. In figure 3 we can see an extract of proximal network representing the project Arabidopsis with the weight value relation computed by our proximal network between the node arabidopsis seen at the centre of the graph and all other nodes which are connected by means of a relation to the node arabidopsis.

Currently, we have successfully processed around 5423 words computing their actual physical occurrence. We have been able to successfully build a proximal network of 50,000 word pair. The documents processed are relating to the research activities carried out in the chosen 3 fields but not limited to

- Arabidopsis thaliana,
- Alteration and reparation of DNA and
- Methodology and speciation

We have constructed different sets of proximal network for each of these domains and the documents used are sub sectioned on the basis of the domain of publication. The documents are then later treated for acronyms and symbols of the domain in question. This information is added to the specific nodes as synonyms after the network is completed.

This data processing method in itself can be independently used for processing and representing data in various domains. The small time taken for processing huge amounts of data makes it an important aspect in the filed of automated ontology construction representing multiple domain scalable.

3.2 Semantic Network

Technically a semantic network is a node- and edge-labelled directed graph, and it is frequently depicted that way. The scope of the semantic network is very broad, allowing semantic categorization [Maedche and Staab, 2001] of a wide range of terminology in multiple domains. Major groupings of semantic types include organisms, anatomical structures, biologic function, chemicals, events, physical objects, and concepts or ideas.

The links between the semantic types provide the structure for the network and represent important relationships. Our semantic network is based on the KL-ONE model [Brachman et al., 1991], with domain being the centre of our network which is expatiated by the domain components which in turn define concepts using the instance and inheritance relations.

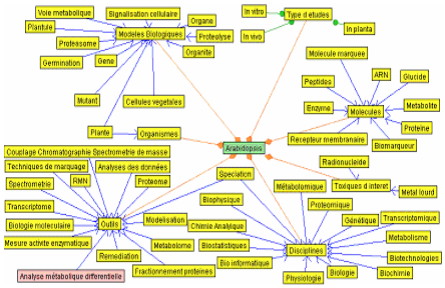

Figure 4: An extract of semantic representation of concepts for arabidopsis using graph editor

We follow the scheme process [Belkin and Croft, 1992] where minimum required information on a domain is precisely represented using the semantic relations defined above. The model is built based on the same set of documents used in proximal network and the 50 most important concepts from the data of these documents are chosen, with the help of a domain expert and are then developed into the semantic model. Here each relational link used namely the compositional, instantiation and inheritance links are given a predefined unit for calculation at a later phase when the network is put into experimentation. This is a precision

model where great emphasis is laid on the semantic relation the nodes share.

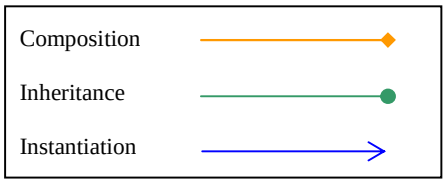

Figure 5: Relational links used in semantic model

The relational links used here are derived from the UML relational links based on the Minsky model of class representation and the KL-One knowledge representation model. The main semantic relations of instantiation and inheritance are retained in our semantic model. The domain is first defined as composed of seven sub models as shown in figure 4. These sub models will remain common to all the domains treated here. We also retain the feature of multiple inheritances in our model on similar argument as in the KL-one model.

The value for the above links are so chosen that every node in a semantic network should always hold a value equal or greater then 50 as a weight on a scale ranging from 0 to 100. This is for the simple reason that in our ESN model the semantic network forms the guiding model and forms the main classes for all the sub classes and roles that are to be added to it from the proximal network.

3.3 Integration of Models

The 2 different resulting models obtained by our methods dilated in the previous sections are combined employing the simple extension methods. We retain our semantic network model as the basic architecture of our knowledge representation model. Here the common nodes between the two networks are identified by our algorithm and this is used to expand the semantic network model using results from the proximal network model. Here the extension is defined based on the proximity value shared by the connected words.

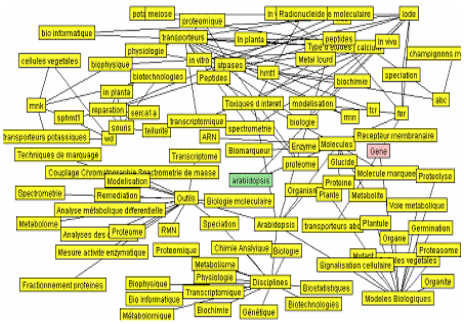

Figure 6: An Extract of Extended Semantic Network Visualised using Graph Editor

At present, we have limited this to a level of 5 extensions i.e. only the first 5 level of nodes are added from the proximal network. The directional flow of the process is restrained where the relational flow is possible only from a lower level node to the upper level node. This model retains all redundant data appearing in different forms when combining the nodes from proximal network.

We begin with the nodes of the semantic network. These nodes in turn guide our algorithms to identify the extension nodes from the proximal network. Thus we elaborate the small semantic network into a large word network with the help of the machine calculated proximal network. This in fact helps in making the right information appear in many forms and consequently correct answers can be filtered from false positives by relying on the correct answers to appear more number of times in a document.

This can be very interesting in a QA retrieval technique where data redundancy in massive collections of documents plays an important role in determining the specific information. We are also verifying other possible methods of merging the 2 networks. One of the most interesting methods is considering proximal network to be a source network providing roles to the semantic network based on the guidelines of KL-One model.

4 Results and Future Work

The Extended Semantic Network prototype has been developed in collaboration with the ToxNuc-E project funded by CEA (Commissariat à l'Energie Atomique). ToxNuc-E [Ménager, 2004], is a project devoted to all the research activities carried out for controlling nuclear environmental toxicology in the living environment with several research centres like CNRS, INSERM etc involved. It is a platform where researchers from different domains like biology, chemical, physics and nuclear, across Europe working for a common purpose, meet and exchange their views on various on-going research activities related to nuclear toxicology.

The ToxNuc-E presently has around 660 researchers registered with their profile, background and area of research interest geographically displaced. Our research is applied in this platform to provide these researchers knowledge representation tool like ESN which can be utilized in information retrieval problems of finding limited and specific information. Currently, we are experimenting on the 3 topics chosen by the researchers as the domain of major research activities. The data and the documents used in our experimental prototype of ESN are obtained from the ToxNuc-E platform. We soon intend to extend our research to all of the research fields of ToxNuc-E.

The results of our algorithm have been subjected to testing, by human experts and have been judged to provide results very close to human constructed concept networks [Rosch, 1978] with reduced time of construction and has proved to be very cost effective. Our qualitatively measured formative results show that ESN is several times faster with a high recall percentage then a human constructed network.

Another important feature of ESN is its ability to customise to user needs and equally provide results very close

to NLP-based indexing methods without much heavy computations i.e. if a user needs specific information on specific subject it is adequate to change the input documents for the proximal network. Based on these documents the entire network is reconstructed in a time span of 30 minutes. This network can then be combined with the existing semantic network. This merged network when used will provide added relevant information on the topic in question.

As an application of our model a document classifier has been developed and integrated on the Toxnuc-E platform. This document classifier uses the ESN knowledge model to classify documents based on their inclination to the 15 research projects that are been piloted by ToxNuc-E. For Initial testing we selected a set of new publications from the project Arabidopsis and MSBE. These documents were then classified using ESN model. The classifier provided an output with the domain inclination percentage.

The same documents were then manually classified by the researchers. We noticed that the results by our classifier highlighted information about certain documents belonging to the original domain Arabidopsis showed inclination to other domains like MSBE a detail not specified until and unless the document is completely read by the user. This information was seen missed by the manually classified result. The correctness of our classifier results were verified by the domain experts who manually verified the documents and confirmed its inclination to both the projects.

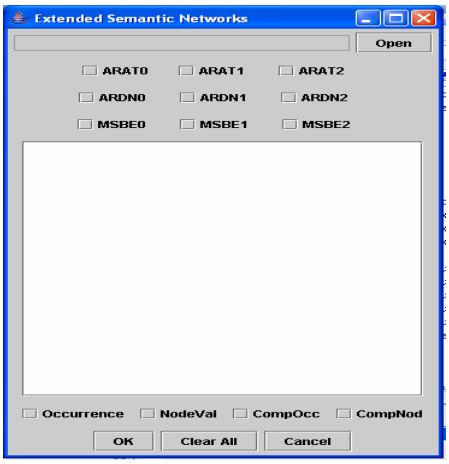

Figure 7: Document classifier

The principle advantage of our methodology with respect to the previous work is our innovative hybrid approach of integrating machine calculations representing our proximal network with human reasoning abilities representing the semantic network built by experts. We use the precise, non estimated results provided by human expertise in case of semantic network and merge them with the machine calculated knowledge network from proximal results. We are now concentrating on developing a search tool based on the KR model constructed by us and intend to publish the benchmark results of our proposal.

5 Acknowledgement

We would like to use this opportunity to thank all the researchers of ToxNuc-E platform for providing us with their data and useful expertise. We would also like to thank the reviewers of this paper for their useful comments.

6 References

[Moldovan, 2002] D. Moldovan, M. Pasca, S. Harabagiu and M. Surdeanu, “Performance issues and error analysis in an open-domain question answering system”, in Proceedings of ACL 2002.

[Voorhees, 2002] Ellen Voorhees. Overview of the TREC 2002 Question Answering Track. In Trec 2002.

[Katz, 1997] Boris Katz. Annotating the world wide web using Natural Language. In Proceedings of the 5th RIAO conference on computer Assisted Information Searching on the internet, 1997.

[Sowa, 2000] J.F Sowa, Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA, 2000.

[Quillian, 1968] M.R Quillian,, Semantic memory. M Minsky, Ed, Semantic Information Processing. pp.216-270. Cambridge, Massachusetts: MIT Press, 1968.

[Sowa, 1984] J.F Sowa, Conceptual structures: information processing in mind and machine, Addison-Wesley Longman Publishing Co., Inc, Boston, MA, 1984.

[Brachman et al., 1991] J Brachman, L Deborah, McGuinness, F Patel-Schneider, A Resnick Living with CLASSIC: When and How to Use a KL-ONE-Like Language, Special issue on implemented knowledge representation and reasoning systems Pages: 108 – 113, ACM Press, NY, USA,1991.

[Brachman and Schmolze, 1985] Ronald J. Brachman and James G. Schmolze. An overview of the KL-ONE knowledge representation system. *Cognitive Science*, 9(2):171–216, April-June 1985.

[Shetty et al., 2006] Reena T N Shetty, Pierre M Riccio and Joël Quinqueton. Hybrid method for knowledge processing, integration and representation. IEEE-IRI 2006 proceedings, September 2006.

[Shetty et al., 2006] Reena T. N. Shetty, Pierre M. Riccio and Joël Quinqueton. Extended semantic network for efficient knowledge representation- An hybrid model. IFIP series by Springer, September 2006.

[Winston et al., 1987] M.E Winston, R Chaffin and D Hernnann, A taxonomy of part – Whole Relations Cognitive Science 11, 1987.

[Mahé et al., 2001] S.A. Mahé, P.M. Riccio et S. Vailliès: des elements pour un modèle: la lutte des classes! Revue Génie Logiciel, n°58, Paris, septembre 2001.

[Ménager, 2004] M Ménager, Programme Toxicologie Nucléaire Environnementale : Comment fédérer et créer une communauté scientifique autour d’un enjeu de

société , Intelligence Collective Partage et Redistribution des Savoirs, Nimes, France, septembre, 2004.

[Maedche and Staab, 2001] Alexander maedche & Steffen Staab, “Ontology Learning for the Semantic Web”, Volume 16 IEEE Intelligent Systems, 2001

[Rosch, 1978] E Rosch Cognitive Representation of Semantic Categories, University of California, Berkeley, 1978

[Belkin and Croft, 1992] N.J Belkin, W.B Croft, Information Filtering and Information Retrieval: Two Sides of the Same Coin? Communications of the ACM Vol. 35 n°12, 1992